Response to Referee #2

Manuscript: egusphere-2026-969

Shifts in the Dominance of Climatic, Anthropogenic, and Landscape Drivers Explain the Spatial Variation in the Budyko-type Equation Parameter

Reviewer's general comment

The reviewer recognizes the potential value of the framework but raises concerns about methodological clarity, validation, sample size, interpretation of model-derived responses, and the strength of the hydrological claims.

Response: We sincerely thank the reviewer for the detailed and technically thoughtful assessment. We appreciate the recognition that the framework has the potential to make a valuable contribution. We also understand that the central concern is whether the analytical design and its presentation provide sufficient support for the level of interpretation used in the original manuscript. We respectfully believe that these concerns can be addressed through a focused and verifiable revision without changing the central research question.

The study contains two distinct analytical levels that were not separated clearly enough in the original presentation. Functional zoning is based on the long-term characteristics of 12 catchments, whereas the temporal dataset contains 31 moving-window observations for each catchment. These windows provide repeated temporal information, but they do not create 372 independent spatial units. For spatial generalization, the effective number of independent catchment units remains 12, and the overlap among adjacent 11-year windows further requires a validation design that explicitly respects both temporal and catchment grouping.

Following the review, we re-examined the data-processing and zoning workflow. We confirmed the population-density spreadsheet error identified by the reviewer and traced it to the final density conversion for one group rather than to the underlying gridded population totals. After correcting this conversion and rechecking the remaining predictors, our preliminary corrected-input analysis recovers the same three-zone membership. This initial check indicates that the grouping is not driven by the population-density error. Nevertheless, the PCA, clustering robustness, and all downstream model results will be recalculated before any final claim about zone stability is made.

In the revision, we will replace the random validation scheme with leakage-resistant temporal and catchment-based validation. The primary model comparison will use all 12 catchments, and strictly out-of-fold predictions and SHAP results will then be summarized by functional zone. Separate within-zone fits will be retained only as sensitivity analyses because some zones contain as few as three catchments. We will also report the effective sample structure, provide the complete PCA and clustering diagnostics, and distinguish model-relative SHAP response changes from universal physical thresholds. We will retain hydrological interpretation where it is supported, but will remove causal or mechanistic statements that cannot be separated from confounding. These actions are described point by point below.

Comment 1. Dependence between moving windows and potential train-test leakage

The reviewer notes that consecutive 11-year moving windows share approximately 10 of 11 underlying years. A random 70/30 split can therefore place near-duplicate windows in the training and test sets, which may inflate the reported R^2 values in Table 2 and Figs. 5-6. The reviewer also argues that the leave-one-basin-out results should be part of the main evaluation rather than a brief limitation.

Response: This is an important concern, and we agree that the original random split does not provide a sufficiently conservative estimate of transfer performance for overlapping moving-window data. The overlap does not affect the catchment-level zoning matrix, but it can inflate the apparent predictive performance of the original models.

We will therefore replace the original random split as the basis for model assessment. The primary spatial evaluation will pool the modelling data from all 12 catchments and use leave-one-catchment-out validation. All windows from one catchment will be held out together, and the model will be trained on the remaining 11 catchments. The held-out predictions and SHAP values will then be summarized by the three functional zones. The zone label will be used for stratified interpretation rather than as a predictor. This design preserves the study's zonal comparison while avoiding the unstable situation in which a three-catchment zone is trained on only two catchments. Separate within-zone models will be presented only as an exploratory sensitivity analysis and will not provide the primary performance estimates.

A complementary temporally blocked evaluation will separate earlier and later periods. Because each window spans 11 years, adjacent training and test blocks will be separated by a purge of at least 10 window positions so that no underlying calendar year occurs in both sets. All preprocessing, feature selection, and hyperparameter tuning will be performed within the training data and will respect the corresponding catchment or temporal grouping.

We will report R^2, RMSE, and MAE from these leakage-resistant evaluations in the main Results and use them, rather than the original random-split values, when comparing algorithms and discussing predictive skill. Uncertainty will be summarized across catchments rather than treating the overlapping windows as independent replicates. If the revised performance is lower, the Abstract, Results, Discussion, and Conclusions will be adjusted accordingly. This directly addresses the related annotations on the apparently high R^2 values and the validation samples shown in Figs. 5 and 6.

Comment 2. Limited number of catchments and precision of the reported response values

The reviewer questions whether 12 catchments, and groups containing only 3-5 catchments, can support generalized statements for the semi-arid to semi-humid transition zone. The reviewer recommends either increasing the number of catchments or reframing the study as a regional case study, and suggests reporting ranges rather than highly precise SHAP-based point values.

Response: We agree that 12 catchments do not support population-level generalization to all transition regions. However, the purpose of the catchment set is a regional comparative diagnosis of spatially contrasting hydrological behavior, not estimation of a universal frequency distribution of catchment types. We will make this scope explicit and frame the work as a multi-catchment case study in northern China. Claims extending to semi-arid and semi-humid transition zones in general will be removed or rewritten as hypotheses for testing in larger samples.

The revised manuscript will report both levels of sample size. The zoning analysis uses $n = 12$ catchments and $p = 13$ long-term catchment attributes. The temporal modelling dataset contains 31 windows per catchment, but these windows will not be described as fully independent observations. Exact catchment and validation-sample counts will accompany every model-performance figure.

We will also replace overly precise single response values with zone-stratified, out-of-fold response summaries from the pooled leave-one-catchment-out analysis. Catchment-level sensitivity will be evaluated without treating the 31 overlapping windows as independent replicates. A response transition will be retained only when its direction and approximate location are stable across held-out catchments and under the temporally blocked analysis. Unstable features will be described without a numerical transition value. This preserves the regional diagnostic contribution while aligning the precision of the claims with the available catchment evidence.

Comment 3. Definition and justification of the hydroclimatic transition region

The reviewer asks for the source and method underlying the climatic classification in Fig. 1, a catchment-level climatology table, and a clearer justification of why catchments located within one climatic category are included. The reviewer also raises the possibility that the functional zoning merely reproduces the preselected climatic pattern.

Response: We agree that the original Study Area section did not provide enough information to evaluate the climatic context of the catchments. We will identify and cite the exact source and classification rule used for the background zones in Fig. 1. We will also add a catchment-level table reporting location, drainage area, long-term precipitation, potential evapotranspiration, temperature, aridity index, and climatic category. The approximate spatial extent of the 12 catchments, 101.6-118.0° E and 33.7-40.0° N, will be added to the Study Area description.

We will clarify that 'transition region' refers to the regional transect formed collectively by catchments distributed along the semi-arid to semi-humid gradient. It does not imply that every individual catchment straddles a mapped climatic boundary. Wording such as 'a vast climatic gradient' will be replaced with a quantitative description of the observed climatic range.

The mapped climate-zone label was not used as an input to the PCA or clustering. Nevertheless, we agree that including hydroclimatic attributes and selecting catchments along a climatic transect can cause the resulting groups to reproduce the pre-existing climate gradient. We will therefore compare the full clustering with a climate-only baseline and with a sensitivity analysis that excludes the hydroclimatic predictor block. We will report the complete loading structure and catchment co-membership across these analyses. If the grouping is not distinguishable from the climatic baseline, we will describe it as a descriptive stratification of the regional transect rather than as a newly identified functional classification.

Comment 4. Analytical derivation of ω and structural dependence on climatic predictors

The reviewer notes that ω is inferred from P, R, and E0 rather than directly observed. The reviewer asks whether unmodelled processes are absorbed into the target and whether using P and AI as predictors partly reconstructs the Fu equation rather than identifying independent drivers. The annotated PDF also asks how one ω value is obtained for each 11-year window.

Response: The reviewer is correct that ω is an effective parameter inferred from the long-term water balance. For each 11-year window, annual P, R, and E0 are aggregated to their mean-annual values and the Fu equation is inverted once to obtain one ω value for that window. We will state this calculation explicitly and distinguish inferred ω from an independently observed target.

This effective ω can integrate vegetation, soil, topography, climatic seasonality, storage, reservoir regulation, and other processes not represented explicitly. It should therefore not be interpreted as a landscape-only constant. We will revise the conceptual framing accordingly.

We also agree that P and AI are structurally connected to the inferred target. This is intrinsic to the Fu formulation: P and E0 define the hydroclimatic condition, and P, R, and E0 are used to infer ω. It does not make climatic predictors scientifically inadmissible, because effective ω has been linked to aridity, precipitation depth, and climate seasonality (e.g., Ning et al., 2017; Saha et al., 2020; Cheng et al., 2022). We will retain P and AI in the full diagnostic model, while interpreting their SHAP importance as potentially reflecting both physical hydroclimatic dependence and mathematical coupling, not independent causation.

We will quantify this coupling under leakage-resistant validation by comparing the full model, a Fu-linked climatic baseline, and sensitivity models excluding P and AI separately and jointly. These ablations do not alter the calculation of ω. They test whether anthropogenic and landscape predictors add out-of-fold skill and retain stable rankings beyond the climatic baseline. Because the target remains derived from P, R, and E0, all interpretations will remain diagnostic and associative rather than causal.

Comment 5. Reservoir regulation and interpretation of GDP density

The reviewer notes that reservoir and dam regulation are not represented explicitly and may have major implications for ω. The reviewer further argues that GDP density may covary with hydropower and reservoir infrastructure, which cannot be distinguished from the proposed ecological-investment mechanism.

Response: We agree that reservoir regulation is a potentially important confounder. Because ω is inferred from runoff as well as climate, reservoir operations can influence the inferred target directly, even when reservoir characteristics are absent from the predictor set. The original interpretation of GDP density as evidence of ecological investment was therefore too specific.

We will remove that mechanistic attribution. GDP density will be described only as a composite indicator associated with socioeconomic activity and infrastructure intensity. We will compile available catchment-level information on upstream reservoirs and regulation intensity. Where the information is sufficiently complete, a regulation descriptor will be included in the predictor or sensitivity analysis. At minimum, we will identify strongly regulated catchments and repeat the main comparisons with those catchments flagged or excluded. Reservoir regulation, groundwater abstraction, and other unrepresented management processes will remain explicit competing explanations. Consequently, the GDP response will not be presented as evidence that economic development causes a particular hydrological change. This correction addresses both the C3 interpretation and the related annotated comment on the predictor table.

Comment 6. Transparency and robustness of PCA and hierarchical clustering

The reviewer asks for the dimensions and construction of the PCA matrix, the component-retention rule, silhouette results across cluster numbers, the motivation for clustering, and an assessment of robustness given $n = 12$ and $p = 13$. The reviewer also asks whether the mixed climatic and development loadings on PC1 complicate attribution.

Response: We agree that the original Methods did not provide enough information to reproduce the zoning. The revised description will specify that the PCA input is a 12×13 matrix. Rows represent catchments and columns represent the long-term means of 13 standardized catchment attributes. All variables are z-score standardized before PCA. We will report the eigenvalues, variance explained by every retained component, the cumulative-variance criterion, the complete loading matrix, Ward linkage, Euclidean distance, and the tested range of cluster numbers. Silhouette scores and the dendrogram will be presented rather than reporting only the selected solution.

We acknowledge that n is small relative to p. PCA and clustering will therefore be framed as exploratory dimension reduction and regional grouping, not as population-level statistical inference. Robustness will be examined by repeating the analysis after excluding one catchment at a time and after removing highly correlated predictors or predictor groups. We will report how often catchment pairs remain in the same cluster. Our preliminary corrected-input analysis, after correcting population density and rechecking the remaining predictors, retains the same three-zone membership. This preliminary result will be reported as an initial stability check, while the final choice of k and its uncertainty will be documented transparently.

The coexistence of climatic and development loadings on PC1 does not by itself indicate a failure of PCA. PCA identifies dominant covariance gradients and is not designed to force climatic and anthropogenic variables onto separate axes. We will nevertheless avoid assigning a single causal meaning to PC1 and will interpret the loading directions and PC2 jointly.

Comment 7. Sample sizes and presentation of Figs. 5 and 6

The reviewer asks whether the violin distributions in Fig. 5 contain only 3-5 catchment-level scores and what observations are shown in Fig. 6. The reviewer considers the current plots and captions insufficiently transparent.

Response: The reviewer is correct that the violin plots in Fig. 5 summarize only the catchment-level scores available within each zone. A kernel-density shape is not informative with such small n and can visually imply more distributional information than is present. We will replace the violins with individual catchment points and a simple median marker, without a kernel-density estimate. Every panel will display the number of catchments explicitly.

Figure 6 will be regenerated from the pooled leave-one-catchment-out predictions described in Response 1 and displayed by functional zone. Its caption will state whether each point represents a held-out catchment-window prediction or a catchment-level summary, the number of catchments and predictions in each zone, and the temporal period represented. The revised figures will therefore show the real validation structure rather than pooling observations without a visible sampling hierarchy.

Comment 8. Physical interpretation and stability of the SHAP response changes

The reviewer questions whether the temperature, LAI, and GDP response values are statistically and physically defensible. In particular, the reviewer interprets the sign changes near T = 8.17 °C and LAI = 1.16 as contradictory to standard hydrological expectations and notes that the proposed mechanisms are weakly supported.

Response: We appreciate this comment because it exposes an important ambiguity in our original use of the word 'threshold'. A SHAP value is measured relative to the model's expected prediction. A negative SHAP value below 8.17 °C does not mean that an incremental increase in temperature necessarily decreases ω. It means that observations in that temperature range contribute to predictions below the model baseline after accounting for the fitted feature interactions. Likewise, a zero crossing is not by itself evidence of a physical regime shift. The shape and slope of the dependence curve, its interaction structure, and its stability across catchments must be considered.

We will therefore replace 'critical threshold' and 'tipping point' with 'model-identified response transition' unless independent evidence supports a physical threshold. SHAP values for held-out catchments will be computed from the corresponding pooled leave-one-catchment-out models and then summarized by zone using a consistent training-fold background procedure. The resulting response ranges will be compared with the temporally blocked analysis. If the response direction or zero-crossing location is not stable, no numerical transition will be reported and the associated mechanistic interpretation will be removed.

For LAI, we will remove the unsupported explanations involving soil saturation and agricultural drainage unless independent catchment evidence can be provided. For GDP, both the low- and high-value segments will be described as model associations potentially confounded by reservoir and infrastructure intensity. For temperature, we will distinguish the sign of the SHAP contribution from the marginal direction of the fitted response and discuss alternative explanations only where they are supported by the data.

Comment 9. Causal and mechanistic language

The reviewer notes that the cluster descriptions and later interpretations use causal or mechanistic phrases before the analyses establish those mechanisms.

Response: We agree that several phrases in the original manuscript exceeded what can be inferred from an observational, model-interpretation analysis. We will separate descriptive results from mechanistic hypotheses. In the Results, terms such as 'shaped by', 'driven by', 'physical basis', and 'causes' will be replaced with calibrated language such as 'characterized by', 'associated with', or 'contributes to the fitted prediction'. Potential mechanisms will be discussed only in the Discussion and clearly labelled as interpretations or hypotheses.

This change does not remove the observed spatial contrasts or the model-identified associations. It aligns the strength of the language with the evidence and prevents SHAP attribution from being read as causal identification.

Comment 10. Implausible population density

The reviewer questions the reported population density of approximately 1.1 × 10^7 persons km^{-2} and asks whether the values were checked.

Response: The reviewer is correct that this value is impossible. We rechecked the complete population processing chain. The gridded GlobPOP Count product was correctly summed to obtain basin population totals. The error occurred in the final spreadsheet conversion for one group, where the population total was divided by the cropland-proportion column rather than by basin area.

Using the correct basin-area denominator gives long-term mean population densities of approximately 687.8, 267.9, and 460.3 persons km^{-2} for the three affected catchments, rather than values on the order of 10^7. We will replace the erroneous values throughout the data table, PCA input, Results, figures, and supplementary materials, and rerun all downstream analyses. A preliminary recalculation shows that this correction does not change the membership of the three functional zones. We thank the reviewer for identifying this error.

Comment 11. Framing ω as a landscape-only parameter

The reviewer notes that ω may also contain information on climatic seasonality and cites Yokoo et al. (2008) as an example, questioning the landscape-only framing.

Response: We agree that describing ω as a landscape-only parameter is too restrictive. In the Fu framework, the inferred parameter represents the integrated effect of catchment characteristics and processes not represented by the mean aridity index alone. These may include vegetation, soils, topography, intra-annual climate seasonality, storage dynamics, and human regulation. We will revise the Introduction and conceptual framework to use 'effective catchment parameter' and will add the relevant seasonality literature, including the study suggested by the reviewer.

Comment 12. Abstract, study-area, and framework clarity in the annotated PDF

The annotated PDF suggests adding the geographic extent, defining C1-C3 before first use, improving the analytical-framework wording, and making clear which analyses use 12 catchment means and which use catchment-window observations.

Response: We thank the reviewer for these presentation suggestions. The revised Abstract will introduce the three zone labels before reporting zone-specific findings and will avoid presenting unexplained C1, C2, and C3 labels. The Study Area will report the geographic extent and catchment climatology. The analytical-framework paragraph and Fig. 2 caption will be rewritten to distinguish three data objects explicitly: the 12 × 13 catchment-level zoning matrix, the 31 moving windows per catchment used for temporal modelling, and the held-out observations used for validation. This clarification will also resolve the ambiguity about whether the whole analysis is based on only 12 values.

Comment 13. Code and data availability

The reviewer notes that no code-availability statement is provided despite the complexity of the machine-learning and SHAP workflow.

Response: We agree that computational reproducibility is essential for this study. If invited to submit a revised manuscript, we will provide a Code and Data Availability statement and archive the processing, PCA and clustering, validation, model-fitting, and SHAP-analysis scripts in a persistent repository. The archive will include the processed basin-level input tables required to reproduce the figures and results, subject to the redistribution terms of the original gridded datasets. Repository access and a persistent identifier will be included with the revised submission.

Comment 14. Additional wording and presentation annotations

The annotated PDF identifies several passages that are difficult to read or potentially too strong, including wording about a 'vast climatic gradient' and descriptions that may imply the functional zones are established physical entities rather than data-derived groupings.

Response: We will conduct a targeted terminology and readability revision across the Abstract, analytical framework, Results, Discussion, figure captions, and Conclusions. Qualitative superlatives will be replaced with quantitative ranges. Functional zones will be introduced as data-derived regional groupings whose hydrological interpretation is supported, but not proven, by the observed attribute contrasts. Sentences highlighted as difficult to read will be rewritten for direct subject-verb structure. These presentation changes will remain consistent with the methodological and claim-scope revisions described above.

We again thank the reviewer for the rigorous and constructive comments. The final-response phase is intended to describe the actions that will be taken if a revised submission is invited. We will implement the analyses and revisions specified above and provide a traceable point-by-point account of the resulting changes.